Do's and Don'ts of Computer-Supported Content Analysis: Good reference guide

Robert (Hogenraad, Emeritus) and Dean's (McKenzie, Monash U.) Preferences

robert.hogenraad@uclouvain.be

For the PROTAN software of computer-aided text analysis,
see a presentation note archived at https://archive.org/details/presentation_201711).

Table of contents to the "Good Reference Guide in Do's and Don'ts of Computer-Supported Content Analysis")

1.The very act of handling texts distributed over time entails consequences

- A. We will do well to remind ourselves that when one looks at data, textual or not, over time, there is almost always change
- B. We will do well to remind ourselves that much of the data encountered in psychological research are not normally distributed. Normal distribution is very rare when it comes to real social scientific data.
- C. The issue of the amount of change that takes place between the beginning and end of a series of texts may cause us to suspect biasing effects due to the presence of serial dependencies (autocorrelations) in the texts.
- D. Looking at texts over time also invites one to the next question of where does change occur in the text.
- E. Everything correlates with everything else. Sometimes.

2. Comparisons between correlations can now be performed. Meng et al. show it.

3. Literature has no competitor (or what to do when $N = 1$)

PART I. Creating Meaning Values with Content Analysis: Strategies

Weber, R. P. (1983). Measurement models for content analysis. *Quality and Quantity*, *17*, 127-149. [A general and clear exposé of available strategies].

1. Value-Added Strategies for content analysis.

Not any text should be content-analyzed using any method. The more a text is structured, the more the method allows for unstructured strategies.

Spence, D. P. (1982). *Narrative truth and historical truth. Meaning and interpretation in psychoanalysis*. New York: W. W. Norton.

2. Pressure towards novelty as governing drive behind a text

"Every great poet must inevitably innovate upon the example of his predecessors" (Shelley, "A defense of poetry", p. 169). A serviceable concept to understand what makes a text change is pressure towards novelty. "What can I do that has not yet been done" is the question that any writer (poet, scientist, ...or political figure) has to answer before starting to write. Colin Martindale has expounded on the notion and demonstrated its efficacy.

Holt, R. R. (1956). Gauging primary and secondary processes in Rorschach responses. *Journal of Projective Techniques*, *20*, 14-25.

Shelley, P. B. (1821/1970). A defence of poetry. In R. A. Duerksen (Ed.), *Percy Bysshe Shelley "Political writings"*, including "A defence of poetry" (pp. 164-197). New York: Appleton-Century-Crofts.

Martindale, C. (1975). R*omantic progression: The psychology of literary history.* Washington, DC: Hemisphere.

Martindale, C. (1979). The night journey: Trends in the content of narratives symbolizing alteration of consciousness. *Journal of Altered States of Consciousness*, *4*, 321-343.

Martindale, C. (Ed.). (1988). *Psychological approaches to the study of literary narratives.* Hamburg: Helmut Buske Verlag.

Martindale, C. (1990). *The clockwork muse: The predictability of artistic change.* New York: Basic Books.

3. Can emotional valence in texts be determined from words?

Bestgen, Y. (1994). Can emotional valence in stories be determined from words? *Cognition and Emotion*, *8*(1), 21-36.

Johnson-Laird, P. N., & Oatley, K. (1989). The language of emotions: An analysis of a semantic field. *Cognition and Emotion*, *3*(2), 81-123.

Miall, D. S. (1988). Affect and narrative: A model of response to stories. *Poetics*, *17*, 259-272.

Miall, D. S. (1992). Estimating changes in collocations of key words across a large text: A case study of Coleridge's notebooks. *Computers and the Humanities*, *26*(1), 1-12.

4. Thematics.

There is presently (early 1998) a return of interest for thematics. The following references are not concerned with content analysis, but they add substance to what content analysts have been doing for some time.

Rimmon-Kenan, S. (1995). What is theme and do we get at it? In C. Bremond, J. Landy, & T. Pavel (Eds.), *Thematics: New approaches* (pp. 9-19). Albany: State University of New York Press.

Sollors, W. (Ed.) (1993). *The return of thematic criticism*. (Vol. 18). Cambridge, MA: Harvard University Press.

5. Readability

Daoust, F. (1992). SATO: Système d'analyse de texte par ordinateur (Version 3.6) [Reference manual]. Montréal: Centre d'analyse de textes par ordinateur (ATO), Université du Québec à Montréal. [The SATO software handles much more than mere readability].

Gunning, R. (1968). *The technique of clear writing*. (Revised ed.). New York: McGraw-Hill Book Company.

Mailloux, S. L., Johnson, M. E., Fisher, D. G., & Pettibone, T. J. (1995). How reliable is computerized assessment of readability? *Computers in Nursing*, *13*(5), 221-225.

6. Computer-supported content analysis of psychiatric data.

The Bucci references do not contain content-analytic strategies per se, but they point toward hypotheses that can usefully be tested with the help of imagery and emotion content-analytic dictionaries used in combination.

Bucci, W. S. (1984). Linking words and things: Basic processes and individual variation. *Cognition*, *17*, 137-153.

Bucci, W. S. (1985). Dual coding: A cognitive model for psychoanalytic research. *Journal of the American Psychoanalytic Association*, *33*, 571-607. [Bucci has replaced Freud's zigzag theory of the relationship between word and image (based upon the economy principle) with the dual code theory of Paivio (1971)].

Bucci, W. S. (1988). Converging evidence for emotional structures. In H. Thomae, H. Kaechele, & S. Hahl (Eds.), *Psychoanalytic process research strategies*. New York: Springer.

Bucci, W. S. (1988). A reconstruction of Freud's tally argument: A program for psychoanalytic research. *Psychoanalytic Inquiry*, *9*, 281-302.

Dahl, H. (1972). A quantitative study of a psychoanalysis. In R. R. Holt & E. Peterfreund (Eds.), *Psychoanalysis and contemporary science* (Vol. 1, pp. 237-257). New York: MacMillan.

Dahl, H. (1974). The measurement of meaning in psychoanalysis by computer analysis of verbal contexts. *Journal of the American Psychoanalytic Association*, *22*, 37-57.

Dahl, H. (1983). On the definition and measurement of wishes. In J. Maslinger (Ed.), *Empirical studies of psychoanalytic theories* (Vol. 1, pp. 39-67). Hillsdale, New Jersey: Lawrence Erlbaum. [Computer-aided content analysis of psychoanalytic sessions; the Dahl's 1983 reference is a variation of the 1972 one; talks about correlating words with the Harvard first- and second-order categories -see below the Meng's et al. 1992 reference on comparisons between correlated correlations-].

Dahl, H., Hölzer, M., & Berry, J. W. (1992). *How to classify emotions for psychotherapy research*. Ulm, Germany: Ulmer Textbank. [A classification of emotions using the de Rivera's decision theory of emotions].

Dahl, H., Kächele, H., & Thomä, H. (Eds.). (1988). *Psychoanalytic process research strategies*. New York: Springer-Verlag. [Computer-aided content analysis with the Ulmer Textbank].

Mergenthaler, E., & Kächele, H. (1988). The Ulm Textbank management system: A tool for psychotherapy research. In H. Dahl, H. Kächele, & H. Thomä (Eds.), *Psychoanalytic process research strategies* (pp. 195-211). New York: Springer Verlag.

Reynes, R. L., Martindale, C., & Dahl, H. (1984). Lexical differences between working and resistance sessions in psychoanalytic therapy. *Journal of Clinical Psychology*, *40*, 733-737.

Stinson, C. H., Milbrath, C., Reidbord, S. P., & Bucci, W. (1994). Thematic segmentation of psychotherapy transcripts for convergent analyses. *Psychotherapy*, *31*, 36-47.

7. Web pages of interest for content analysis

* Melissa Alexa:

http://www.zuma-mannheim.de/publications/series/working-papers/97_07abs.htm

[An attempt of synthesis. Alexa's page makes no consideration however for the demands of statistical treatment of content-analytic data].

* Content analysis resources:

http://www.gsu.edu/~wwwcom/content.html

[References to books, articles, and computer-aided content analysis software . The web site for the field].

** WordStat, the best of the two worlds:*

WordStat is a content analysis module for the statistical analysis program SIMSTAT for Windows (runs under Win 3.1, Win95 or Win NT). The beta version of WordStat is available publicly. Information regarding the program and download instructions can be obtained from the following URL:

http://www.simstat.com/wordstat.htm

PART II. Statistics for Content-Analytic Data

Statistical tests applied to social science data assume independence of observations. In many cases, content-analytic data are not independent. After all, the chapters of a book are not independent of each other. By definition, we would dare say...

1. The very act of handling texts distributed over time entails consequences:

A. We will do well to remind ourselves that when one looks at data, textual or not, over time, there is almost always change:

Kenny, D. A., & Campbell, D. T. (1984). Methodological considerations in the analysis of temporal data. In K. J. Gergen & M. M. Gergen (Eds.), *Historical social psychology* (pp. 125-138). Hillsdale, NJ: Lawrence Erlbaum.

However, we should not ignore the warning issued by: Montgomery, D. C., & Peck, E. A. (1982). *Introduction to linear regression analysis*. New York: Wiley. ["Extrapolation with polynomial models can be extremely hazardous ... in general, polynomial models may turn in unanticipated and inappropriate directions, both in interpolation and extrapolation" (pp. 183-184). In other words, we should be cautious about forecasting on the basis of a polynomial trend observed for a particular time-series].

B. *We will do well to remind ourselves that much of the data encountered in psychological research are not normally distributed. Normal distribution is very rare when it comes to real social scientific data.*

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*(1), 156-166.

C. The issue of the amount of change that takes place between the beginning and end of a series of texts may cause us to suspect biasing effects due to the presence of serial dependencies (autocorrelations) in the texts:

Burman, P., Chow, E., & Noland, D. (1994). A cross-validatory method for dependent data. *Biometrika*, *81*, 351-358. [Autocorrelation adversely affects cross-validation].

Crosbie, J. (1993). Interrupted time-series analysis with brief single-subject data. *Journal of Consulting and Clinical Psychology*, *61*(6), 966-974.

Glass, G. V., Willson, V. L., & Gottman, J. M. (1975). *Design and analysis of time series experiments*. Boulder, CO: Colorado Associated University Press. ["Analyzed 95 non-cyclical social-behavioral time series, only 6 of which were 2nd order autocorrelation/autoregressive ones, none were higher"].

Hogenraad, R., McKenzie, D. P., & Martindale, C. (1997). The enemy within: Autocorrelation bias in content analysis of narratives. *Computers and the Humanities*, *30*(6), 433-439.

Judd, C. M., McLelland, G. H., & Culhane, S. E. (1995). Data analysis: Continuing issues in the everyday analysis of psychological data. *Annual Review of Psychology*, *46*, 433-465.

Schilds, E., & de Haan, P. (1993). Characteristics of sentence length in running text. *Literary and Linguistic Computing*, *8*, 20-26.

Simonton, D. K. (1977). Cross-sectional time-series experiments: Some suggested statistical analyses. *Psychological Bulletin*, *84*(3), 489-502.

Simonton, D. K. (1990). *Psychology, science, and history: An introduction to historiometry*. New Haven, CT: Yale University Press. [In the case of positive autocorrelation, the N is not as large as it superficially appears, considering how many cases are "a chip off the old (preceding or contiguous) block" (p. 218)].

D. Looking at texts over time also invites one to the next question of where does change occur in the text:

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1993). *Classification and regression trees*. (Revised from 1984 ed.). London: Chapman & Hall. [The CART strategy].

Brodsky, B. E., & Darkhovsky, B. S. (1993). *Nonparametric methods in change-point problems*. Dordrecht, Holland: Kluwer Academic Press.

Siegel, S., & Castellan, N. J., Jr. (1988). *Nonparametric statistics for the behavioral sciences*. (2nd ed.). New York: McGraw-Hill Book Company. [The section on "change-point tests].

Theodossiou, P. T. (1993). Predicting shifts in the mean of a multivariate time series process: An application in predicting business failures. *Journal of the American Statistical Association*, *88*, 441-449.

E. Everything correlates with everything else. Sometimes.

Standing, L., Sproule, R., & Khouzam, N. (1991). Empirical statistics: IV. Illustrating Meehl's sixth law of soft psychology: Everything correlates with everything. *Psychological Reports*, *69*(1), 123-126. [A useful addition to any study concerned with the problems of significance testing, particularly with regards to correlations].

2. Comparisons between correlations can now be performed. Meng et al. show it.

Meng, X., Rosenthal, R., & Rubin, D. R. (1992). Comparing correlated correlations. *Psychological Bulletin*, *111*, 172-175.

3. Literature has no competitor (or what to do when *N* = 1)

Borello, G. M., & Thompson, B. (1989). A replication bootstrap analysis of the structure underlying perceptions of stereotypic love. *Journal of General Psychology*, *116*(3), 317-327. [An example of bootstrap factor analysis].

Diaconis, P., & Efron, B. (1983, 5 May). Computer-intensive methods in statistics. *Scientific American*, *248*, 96-108. [Computer-intensive methods replace standard statistical assumptions about data with massive calculations. One method, the "bootstrap", has revised many previous estimates of the reliability of scientific inferences].

Efron, B., & Tibshirani, R. (1991, 26 July 1991). Statistical data in the computer age. *Science*, *253*, 390-395.

Hogenraad, R., McKenzie, D. P., & van Peer, W. (1998). Literature has no competitor. *SPIEL*, in press.

Cirincione, C., & Gurrieri, G. A. (1997). Computer-intensive methods in the social sciences. *Social Science Computer Review*, *15*(1), 83-97.

Hogenraad, R., et al. (in preparation). It might have gone another way: A counterfactual analysis of political speeches that buoy the European construction, using the resampling strategy. *Social Science Computer Review.*

Péladeau, N. (1996). Simstat for Windows. User's guide (Version 1.21d, November 1997) [Windows]. Montréal, Canada: Provalis Research. [Resampling statistics for a mere $129. Should be on every hard disk! More at http://www.simstat.com.

Stubbs, M. (1995). Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Language*, *2*, 23-55. [About the problem of statistical analysis of literary data in that most statistical tests assume independence of observation, and yet texts are not independent, not random. In writing a text, authors consciously or unconsciously follow a pattern, which is why autocorrelation exists. Thus the normal distribution is not appropriate for textual data].

Thompson, B. (1988). Program FACSTRAP: A program that computes bootstrap estimates of factor structures. *Educational and Psychological Measurement*, *48*, 681-686. [Bootstrap factor analysis: The concept].

Thompson, B. (1995). Exploring the replicability of a study's results: Bootstrap statistics for the multivariate case. *Educational and Psychological Measurement*, *55*(1), 84-94.

Wallace, D. (1997). SIMSTAT for Windows. *Social Science Computer Review*, *15*(3), 310-312. [Reviews Péladeau's resampling engine].

Contact for information: Robert Hogenraad (robert.hogenraad@uclouvain.be)